Nick Howard
Final Assignment
ADEC 7310

Multiple Linear Regression to Model Home Prices using the Ames Data Set

Housing prices are important to individual home buyers. Buying a home is a once in a lifetime investment and many consumers would like to know what represents the true value of a home. If you ask someone what matters to them when they buy a home, they might say that the price of a home is all about the home's location, the size of the backyard, or the age of the home. The Ames data set collected 79 variables that can help us to explain what really matters when someone makes this decision. I will attempt to use some of these variables to model the price of a home.

Regression analysis is a statistical procedure that allows us to describe relationships among independent and dependent variables. It also allows us to estimate the value of a dependent (response) variable from the observed values of the independent variables. There are many types of regression models but the models that I chose to focus on in my analysis were linear regression models. The linear regression model is used to determine the relationship between a dependent variable and one or more independent variables. The dependent variable must be continuous, and the independent variables can be continuous, discrete, or categorical (Schneider). The first step for determining the relationship between variables is to create a scatterplot. If the data appears to have a linear relationship then this should be confirmed using a Pearson correlation test, the R^2 (coefficient of determination) value, and the P value. When performing this literature review, I found that I should be cautious of confounders. These are variables that are not only associated with the independent variable but also other dependent variables (Schneider). The presence of missing values is another common problem in performing regression analysis of a data set. If an independent or dependent variable is missing from the data set then the entire set of observations is excluded which diminishes the sample size used for regression analysis, Missing values can be dealt with in a number of ways such as selecting the mean value for a subset of observations. An example of this would be to calculate the mean weight of female patients between the ages of 50-55 in a study that is collecting data on a hospitals population of female patients and replace the missing weight of female patients in this age group with the mean weight.

Linear regression is a powerful tool for predicting the prices of homes. In 2001 Rosiers et al. published an article titled Landscaping and House Values: An Empirical Investigation. In this article they used regression techniques to determine the affect of landscaping attributes to a home's total value.

In 2001 Nguyen et al. published a paper on a comparison of artificial neural networks performed and multiple regression analysis. Both techniques were used to determine the value of a single-family home based on the criteria of Mean Absolute Percentage Error and Absolute Percentage Error. The dependent variables contained values for square footage, number of bedrooms, number of baths, and year the property was built. The author found that ANN works better based on the MAPE and APE criteria than MRA when a moderate to large sample size is used,

The Ames Housing data set contains data that describes the sale of property in Ames, Iowa from 2006 to 2010 (Decock). There are 80 variables within this data set that could be useful to a home buyer when they home buyer assessed a home's value. This data set consist of 23 nominal data values, 23 ordinal data value, 14 discrete data values, and 20 continuous data values.

In my model I am going to attempt to predict the dependent variable of sales price of a home based off a subset of independent variables. To do this I would like to first review the data visually to examine the variables and find variables that may be of interest to the home buyer. I will also use the R programming environment to verify my assumptions. To start my analysis I loaded the data and reviewed the data for missing values. I noticed that several variables contained columns with missing values.

I created a subset of variables that I thought would be useful in performing my analysis from the train data set from the following categories.

- Home Type
- Location
- Square Footage
- When the home was sold.
- Condition of the home at the time of sale

After I created the subset, I proceeded to review the mean sales price by zoning classification, building type, and external condition of the home to see the variation of means. While this analysis was interesting, I decided not to include these variables in my model.

I explored the relationship between year sold and the sale price of the home. It was interesting to se that a high number of homes were sold at the lower end of the price range during 2006 and 2007. The amount of homes sold dropped each year from 2006 to 2010. I then reviewed the sales price data and found that the data was skewed to the right.

Next, I plotted the numerical independent variables with SalePrice to examine if these variables showed a linear relationship. The variables plotted were related to the square footage of areas of the home and all shows a linear relationship.

I then created a subset of variables that I thought would be useful in creating my model. These data points were all numerical data points. I then measured correlation among the numerical variables using the corrplot() function. I saw that there was high correlation between square footage measurements and sales price. I then created scatter plots to examine the relationship between the variables. All of the variables that I tested had a linear relationship with sales price. I created a linear model using my subset and removed the variables with a non-significant p value. I ran this code on the test data and produced predicted values for a home's sale price.

I chose to use linear regression because it is a model that I understand, and it also allowed me to make adjustments if I saw that a data value was not contributing to the model. I stuck to numerical values for this analysis because I was unsure about how to proceed with the categorical or discrete variables. In the future I would like to learn how to turn categorical values into numerical values. I also think that I need to practice selecting my variables. I would like to see if I could benefit from using backwards, forwards, or stepwise selection of variables. I also would like to learn more about confounders and multicollinearity so that I can avoid the problems that this can present when I run a linear model. Some limitations of this models are that it does not respond well to categorical variables or discrete variables and it only works if there is already a linear relationship between each dependent variable and the independent variable but not relationships among the dependent variables that you select.

I learned a lot about finding a model that is a good fit by using the r squared and p value. I also learned about the hazards of multicollinearity in creating a predictive model and why variable selection is so important. I also learned that there is a limit to using only quantitative variables in your analysis and you can limit your models performance by not taking steps to include categorical and discrete variables.

Works Cited

1) Schneider A, Hommel G, Blettner M. Linear regression analysis: part 14 of a series on evaluation of scientific publications. Dtsch Arztebl Int. 2010;107(44): 776–782. 10.3238/arztebl.2010.0776
2) Des Rosiers François, Thériault Marius, Kestens Yan, and Villeneuve Paul (*2002*) Landscaping and House Values: An Empirical Investigation. Journal of Real Estate Research: 2002, Vol. 23, No. 1-2, pp. 139-162.
3) Nguyen Nghiep and Cripps Al (2001) Predicting Housing Value: A Comparison of Multiple Regression Analysis and Artificial Neural Networks. Journal of Real Estate Research: 2001, Vol. 22, No. 3, pp. 313-336.
4) De Cock (2011) Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project, Journal of Statistics Education Volume 19, Number 3(2011)